



(19)



JAPANESE PATENT OFFICE

## PATENT ABSTRACTS OF JAPAN

(11) Publication number: 09034651 A

(43) Date of publication of application: 07.02.1997

(51) Int. Cl. G06F 3/06

(21) Application number: 07183841

(22) Date of filing: 20.07.1995

(71) Applicant: HITACHI LTD

HITACHI CHUBU SOFTWARE LTD

(72) Inventor: HAYASHI ITSUKI

FUJIOKA YOSHINORI

FUKAYA YASUKATSU

FUKUMOTO SATOSHI

## (54) DISK ARRAY DEVICE

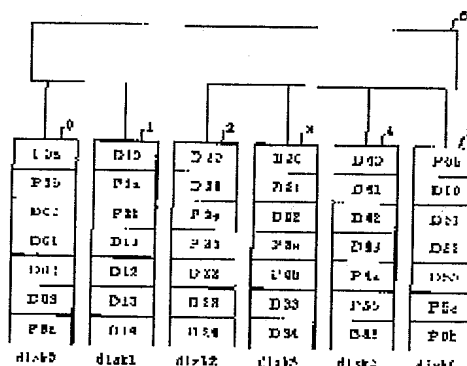
## (57) Abstract:

**PROBLEM TO BE SOLVED:** To reduce disk contention accompanying parity update and to improve the performance of a disk array by forming the same parity group by first and second parities and updating only the first or second parity at the time of parity update.

**SOLUTION:** This device consists of disk devices 0 to 5 (disk 0 to disk 5) of units and an array control mechanism 6. Each of disk devices 0 to 5 is divided into blocks as sets of plural sectors. In this constitution of distributed sparing, spare areas are used as second parity areas, and second parities (POb, P1b...) forming the same parity groups as first parities (POa, P1a...) are stored there. Second parities are provided in the

disk array in this manner, and first and second parities are arranged in the same parity group, and only one of two parities is updated.

COPYRIGHT: (C)1997,JPO



(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平9-34651

(43)公開日 平成9年(1997)2月7日

(51)Int.Cl.<sup>6</sup>

G 0 6 F 3/06

識別記号

5 4 0

庁内整理番号

F I

G 0 6 F 3/06

技術表示箇所

5 4 0

審査請求 未請求 請求項の数7 O L (全 6 頁)

(21)出願番号 特願平7-183841

(22)出願日 平成7年(1995)7月20日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(71)出願人 000233457

日立中部ソフトウェア株式会社

愛知県名古屋市中区栄3丁目10番22号

(72)発明者 林 逸樹

愛知県名古屋市中区栄三丁目10番22号日立

中部ソフトウェア株式会社内

(72)発明者 藤岡 良記

愛知県尾張旭市晴丘町池上1番地株式会社

日立製作所オフィスシステム事業部内

(74)代理人 弁理士 小川 勝男

最終頁に続く

(54)【発明の名称】 ディスクアレイ装置

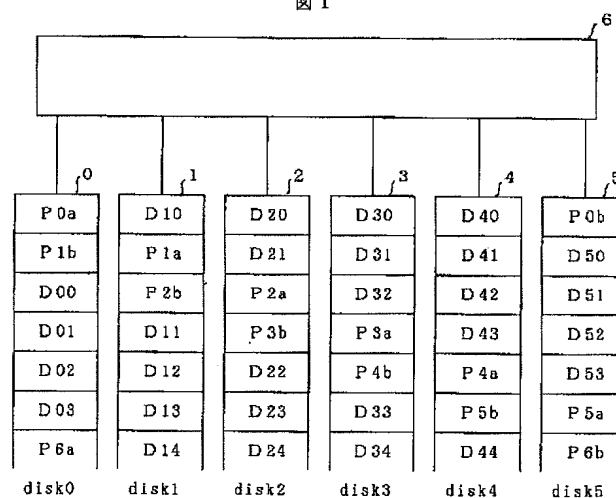
(57)【要約】

【目的】 スペアディスクを常備するディスクアレイの性能の改善を図る。

【構成】 ディスクアレイ内に第1、第2の複数のパリティを設け、第1のパリティおよび第2のパリティを同一のパリティグループに配置し、パリティの更新を2つのパリティのうちの1つに対してのみ行う。障害回復時には、どちらかをスペア領域にする。

【効果】 パリティ競合によるオーバヘッドを軽減でき、ディスクアレイの性能を改善できる。

図 1



## 【特許請求の範囲】

【請求項1】アレイ構成の複数台のディスクドライブとアレイ制御機構からなるディスクアレイ装置であって、該各ディスクドライブは複数のセクタの集まりであるブロックに分割され、データおよびパリティがブロック単位にアレイ内に分散され、各ブロックが独立にアクセスが可能なディスクアレイ装置において、複数のデータおよびパリティとのパリティ計算単位であるパリティグループ内に2以上の複数パリティを含ませたことを特徴とするディスクアレイ装置。

【請求項2】請求項1のディスクアレイ装置において、前記アレイ制御機構に対するライト要求におけるパリティ更新を前記複数のパリティのうちの1つのみを更新することにより実行することを特徴とするディスクアレイ装置。

【請求項3】請求項1または2記載のディスクアレイ装置において、前記複数台のディスクドライブのうちの1台が故障した場合は、前記複数のパリティのうちの1つをスペア領域として、消失データおよび新パリティの生成を行い、リカバリを実行することを特徴とするディスクアレイ装置。

【請求項4】アレイ構成の複数台のディスクドライブとアレイ制御機構からなるディスクアレイ装置であって、該各ディスクドライブは複数のセクタの集まりであるブロックに分割され、各ブロックが独立にアクセスが可能なディスクアレイ装置において、前記複数台のディスクドライブのうちの2以上の複数台をパリティディスクとし、該複数台のパリティディスクのパリティを同一のパリティグループ内に配置したことを特徴とするディスクアレイ装置。

【請求項5】請求項4記載のディスクアレイ装置において、前記アレイ制御機構に対するライト要求におけるパリティ更新を前記複数台のパリティディスクのうちの1つのみを更新することにより実行することを特徴とするディスクアレイ装置。

【請求項6】請求項4又は5記載のディスクアレイ装置において、前記複数台のディスクドライブのうちの1台が故障した場合は、前記複数台のパリティディスクのうちの1つをスペアディスクとして、消失データおよび新パリティの生成を行い、リカバリを実行することを特徴とするディスクアレイ装置。

【請求項7】請求項6記載のディスクアレイ装置において、前記ディスク1台のディスク故障に対するリカバリ動作の後の、新しいパリティディスクを追加して故障前のアレイ構成に戻すコピーバック動作を、前記追加したパリティディスクの内容をオール'0'にイニシャライズすることにより実行することを特徴とするディスクアレイ装置。

## 【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明はディスクアレイ装置に関し、特にライトコマンドに対するパリティ更新のオーバーヘッドを軽減することの可能なディスクアレイ装置に関する。

【0002】

【従来の技術】ディスクアレイは、アレイ構成された複数のディスクを並列アクセスすることによりアクセススピード等の性能の向上を図り、データにパリティを付加して、1台のディスク故障に対して消失データの自動回復が可能な、高性能、高信頼性ディスクシステムである。

【0003】ディスクアレイの基本構成は、Patterson等によりRAID (Redundant Arrays of Inexpensive Disks) として5つのレベルにまとめられ、このうち、レベル4 (RAID4) およびレベル5 (RAID5) は、アクセス単位をブロック単位でインターリーブする方法である。中でもRAID5は、パリティをアレイ内に分散することにより、パリティディスクへの負荷分散を可能とした、最も分散処理システム向けの方法と考えられている。

【0004】RAID5およびRAID4の性能上の最大の問題点は、パリティ更新のために、ライトコマンドが常にリードモディファイライトアクセスになることである。即ち、1回のライトコマンドに対して、4回のディスクアクセス（データおよびパリティのリードとデータおよびパリティのライト）が必要になる。特に、RAID4では、パリティが1台のディスクに集中しているため、パリティを格納しているディスクへのアクセスが性能上のボトルネックになっている。

【0005】このオーバーヘッドの軽減方法として、スペアリングの方法が提案されている。オンライントランザクション処理等のシステムでは、ディスクが故障した場合の回復処理（リカバリ）をシステムを停止することなく実行するために、通常、ホットスタンバイのスペアディスクを常備している。スペアリングとは、スペアディスクの利用技術のことである。Proceedings of 19th Annual International Symposium on Computer Architecture において、Menon等がこの分散スペアリングの方法を提案し、その性能を評価している。

【0006】分散スペアリングは図3に示す如く、各ディスク (disk0～disk5) にパリティ (P0, P1, P2...) だけでなく、スペア領域 (S0, S1, S2...) も分散させる方法である。このスペア領域を利用することにより、データディスクの数を実質的に1台増加させることができ、アレイ性能の改善を図っている。

【0007】

【発明が解決しようとする課題】上記分散スペアリングでは、データディスクの数は増やすことができるが、前述したライトコマンドに対するパリティネックの問題は基本的には改善されていない。即ち、図3において、デ

10

20

30

40

50

ータD20とパリティP0をアクセスするライトコマンド1を実行している時に、データD32とパリティP2をアクセスするライトコマンド2を実行することはできない。

【0008】本発明の目的は、上記問題点を解消し、ライトコマンドに対するパリティの更新に伴うディスク競合を軽減させる方法を提供し、ディスクアレイの性能の改善を図ることにある。更に本方法をRAID4にも適用し、RAID4におけるパリティディスクのボトルネックを大幅に改善することにある。

【0009】

【課題を解決するための手段】本発明の上記目的は、分散スペアリングの構成において、スペア領域(S0, S1, S2...)を、図1に示す如く第2のパリティ領域として使用し、第1のパリティ(P0a, P1a, P2a...)と同一のパリティグループを成す第2のパリティ(P0b, P1b, P2b...)を格納することにより実現できる。また、スペアディスクを常備するRAID4構成のディスクアレイにおいて、スペアディスクを図2に示す如く第2のパリティディスクとして使用し、第1のパリティ(P0a, P1a, P2a...)と同一のパリティグループを成す第2のパリティ(P0b, P1b, P2b...)を格納することにより実現できる。

【0010】

【作用】第1および第2のパリティで同一のパリティグループを形成するため、パリティの更新は第1もしくは第2のどちらかのパリティのみを更新すればよく、パリティ更新に伴うディスク競合を軽減でき、ディスクアレ

$$P0a \text{ XOR } D10 \text{ XOR } D20 \text{ XOR } D30 \text{ XOR } D40 \text{ XOR } P0b = 0$$

の関係を持たせて格納させる。ここで、XORは排他的論理和を表す。

【0017】ディスク0～5(disk0～disk5)の初期化は各パリティグループ毎に行われる。即ち、まず、初期化されたディスクに新たにデータを格納するときには、ディスク装置の横一列のパリティグループについて、4個のデータブロックについて生成したパリティを第1または第2のパリティブロックの一方に書き込み、残りのパリティブロックにオール'0'を書き込んでおく。

【0018】例えば、図1の最上位のパリティグループにおいて、D10, D20, D30, D40について生成したパリティをP0aに書き込み、P0bにはオール'0'を書き込む。それに続くデータについても同様の処理を行なって、データを格納しておく。

【0019】格納されたデータの一部又は全部を書き換える(更新)するときの制御は、図4に示すアルゴリズムに従う。上位システムからアレイ制御機構6にライトコマンドがリクエストされると(処理11)、アレイ制御機構6は、更新するデータの存在するディスク装置と、更新するデータのパリティグループの第1および第2のパリティの存在するディスク装置が動作中(ビジ

\*イの性能改善を図れる。

【0011】また、ディスクの障害回復(リカバリ)時には、第1または第2のパリティをスペア領域として使用するため、ディスクアレイの障害回復機能を維持できる。

【0012】

【実施例】以下、本発明の一実施例を図面を用いて詳細に説明する。

【0013】図1は本発明の一実施例をRAOD5に適用した例を、図2は本発明の一実施例をRAID4に適用した例をそれぞれ示す。図1及び図2において、0～5(disk0～disk5)は単体のディスク装置、6はアレイ制御機構である。

【0014】各ディスク装置0～5はそれぞれ、複数のセクタの集まりであるブロックに分割される。それぞれのブロックには、当該ディスク装置に格納するデータがデータブロック(D01, D02, D03...)と、第1のパリティブロック(P0a, P1a, P2a...)、第2のパリティブロック(P0b, P1b, P2b...)に分割されて前記各ディスク装置のブロックに格納される。

【0015】図4は本発明におけるパリティの更新アルゴリズムの一例を示した図である。

【0016】図1において、第1のパリティと第2のパリティはディスク0～5内に分散されていて、ディスク0～5の横1列のブロックにより、1つのパリティグループが形成される。即ち、パリティを偶数パリティと仮定すると、

一)かどうかをチェックする(処理12, 13)。

【0020】もし書き換えるデータが存在するディスク装置へのアクセスがビジーの場合は、データのディスクのアクセスが終了するまでライトコマンドの実行をウェイトする。そして、書き換えるデータが存在するディスク装置がビジーでない場合には、当該データのパリティが存在するディスク装置がビジーか否かをチェックする。本発明では、書き換える1つのデータブロックに対して2つのパリティが存在するが、このチェックでパリティが存在するディスク装置が両方共ビジーの場合は、どちらかのディスク装置のアクセスが終了するまでライトコマンドの実行をウェイトする。もし片方のパリティの存在するディスク装置がビジーでない場合には、ビジーでないディスク装置のパリティを用いてパリティの更新を行う(処理14)。この際には他方のパリティについてはそのままのパリティとしておく。

【0021】パリティの存在するディスク装置が両方共にビジーでない場合は、あらかじめ決められたどちらかのパリティ(例えば第1のパリティ)を用いてデータブロック及びパリティの更新を行う(処理15)。

【0022】なお、パリティの存在する両方のディスク

装置が共にビジーでない場合に使用するパリティの決め方は、最後のアクセスが古い方を選択する方法などでもよい。

【0023】例として、今、データD20およびパリティP0aへのライトコマンド1を実行中に、データD32へのライトコマンド2がアレイ制御機構6にリクエストされたとすると、アレイ制御機構6は、パリティP2aがビジー中のため、データD32およびパリティP2bでライトコマンド2を実行する。即ち、データD32およびパリティP2bを読みだし、該読みだしたデータおよびパリティとデータD32の更新データにより、更新パリティを生成し、データD32に更新データを書き込むと同時に、パリティP2bに該更新パリティを書き込む。

【0024】図3で示す分散スペアリングでは、パリティP2bがスペア領域であり、パリティP2aがデータD20と競合するため、ライトコマンド2の実行はライトコマンド1が終了するまで待たされることになる。

【0025】一方図2において、第1のパリティ(P0a, P1a, P2a...) および第2のパリティ(P0b, P1b, P2b...) はそれぞれ専用のディスク装置、ディスク装置4 (disk4)、ディスク装置5 (disk5) に格納されている。パリティグループは図1と同様の横1列のブロックで構成される。RAID4では、ディスク装置5 (disk5) がスペアディスクのため、ライトコマンドの並列アクセスが不可能であるが、本構成では、2台のパリティディスク(disk4, disk5)の内の1つのパリティを更新すればよいから、ライトコマンドの並列アクセスが可能になる。

【0026】例えば、データD00およびパリティP0aに対するライトコマンドとデータD11およびパリティP1bに対するライトコマンドの並列アクセスが可能になる。

【0027】図5は図1におけるアレイ制御機構6の内部構成を詳細に示した図である。図5において、0~5はディスク装置、6はアレイ制御機構、7はバスインタフェース、8はリードバッファ、9はアドレス変換機構、10~15はディスクコントローラ、16はマイクロコントローラ、17はマルチプレクサ、18はリードモディファイライト用バッファ、19はXORロジック、20はコマンドバッファ、21はパリティマッピング用テーブルである。上位インタフェース106からのリード/ライトコマンドは、バスインタフェース7を経由して、アドレス変換機構9により、各ディスク装置0~5専用のキューイングバッファであるコマンドバッファ20に割り当てられる。この割り当て方法はマイクロコントローラ16によりプログラマブルに変更できる。

【0028】マイクロコントローラ16はライン108により、コマンドバッファ20のコマンド内容を認識できる。リードコマンドが先頭の場合は、マイクロコントローラ16はライン110により、当該ディスク装置がビジーかどうかを判定し、もしビジーでなければ、当該

ディスクコントローラにリード起動をかけ、当該ディスク装置からデータを読み込む。該読みだされたデータは、リードバッファ8に格納され、ライン107、バスインタフェース7経由で上位インタフェースバス106に送出される。

【0029】一方、ライトコマンドが先頭の場合は、マイクロコントローラ16はパリティマッピング用テーブル21により、該ライトコマンドに対応する2つのパリティのアドレスを調べ、該2つのパリティを持つディスクがビジーかどうかを判定し、ビジーでないパリティを更新する。もし、両方共ビジーでなければ、予め決められたばれてい更新する。

【0030】パリティの更新は以下の手順で行う。マイクロコントローラ16はライトするブロックの古いデータと、当該パリティブロックの古いパリティを、当該ディスク装置より読出し、該読みだした古いデータ及びパリティはマルチプレクサ17経由でリードモディファイライト用バッファ18に格納する。

【0031】次にマイクロコントローラ16はマルチプレクサ17経由でコマンドバッファ20から読みだしたライトデータと、前記リードモディファイライト用バッファ18に格納した古いデータおよびパリティを、XORロジック(排他的論理和)19に入れ、新しいパリティを生成する。該生成した新しいパリティは、ライン109経由で当該ディスクコントローラ経由で、当該ディスク装置に書き込まれる。この時、同時に、ライトデータも当該ディスク装置に書き込まれる。

【0032】上記リード及びライトコマンドは当該ディスク装置が仕様可能であれば、並列に処理することができる。

【0033】次に、ディスク装置に故障が生じた後の回復処理について説明する。スペア領域を用いる図3に示す方式の場合、1台のディスク装置が故障した場合、ディスクアレイ装置の制御装置は消失データの回復処理を実行する。この時、再生成したデータはスペア領域に格納する。本実施例では、第1のパリティ(P0a, P1a, P2a...) または第2のパリティ(P0b, P1b, P2b...) のどちらか一方(所定のルールにより予め決めておく)をスペア領域として使用する。また、残りのパリティは、故障ディスクの替わりの新しいディスクが用意されるまで、通常のRAID5またはRAID4構成で動作させるために再生成をする。消失したパリティの再生成は行わない。

【0034】例えば、図1において、ディスク1(disk1)が故障したとすると、パリティエリアP0bに消失したデータD10を生成し、パリティP0aにはデータD20, D30, D40, D10の新しいパリティを生成して格納する。消失したパリティP1aは再生成せず、P1bにデータD21, D31, D41, D51の新しいパリティを生成する。図2のRAID4においても、同様の処理を行う。

7

【0035】回復処理の終了後、故障ディスクの替わりの新しいディスクが用意されると、ディスクアレイの構成を故障前の状態に戻すコピーバック動作を行なう。

【0036】本動作は、図1に示すRAID5構成の分散パリティでは、新しいディスクにデータをコピーすると共に、2つのパリティを再生成することにより実行する。

【0037】一方、図2のRAID4構成の分散パリティでは、パリティが予め定められた2つの決まったディスクに格納されるので、コピーバック動作は行わない(必要ない)。オール'0'にイニシャライズされた新しいディスクを用意し、第2のパリティディスクとするだけでよい。

【0038】

【発明の効果】以上、説明した如く、本発明によれば、\*

8

\* ディスクアレイ内に第2のパリティを設け、第1のパリティおよび第2のパリティを同一のパリティグループに配置し、パリティの更新を2つのパリティのうちの1つに対してのみ行うことにより、パリティ競合によるオーバヘッドを軽減し、ディスクアレイの性能を改善することができる。

【図面の簡単な説明】

【図1】本発明の第1の実施例を示した図。

【図2】本発明の第2の実施例を示した図。

【図3】従来技術を示した図。

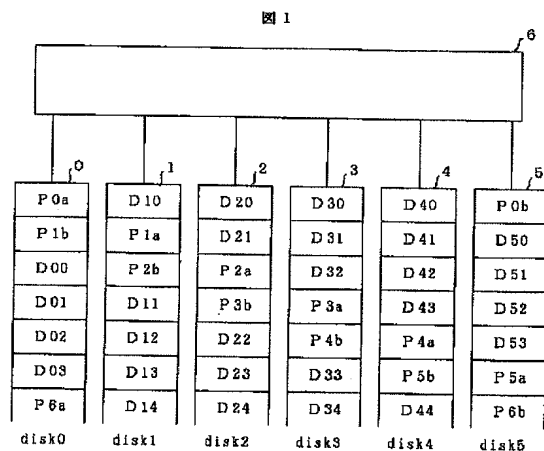
【図4】本発明におけるパリティ更新のアルゴリズムの例を示した図。

【図5】図1の詳細構成を示す図。

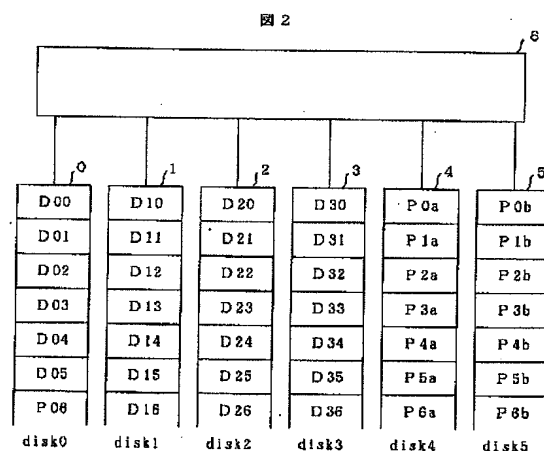
【符号の説明】

0～5…ディスク装置、6…アレイ制御機構。

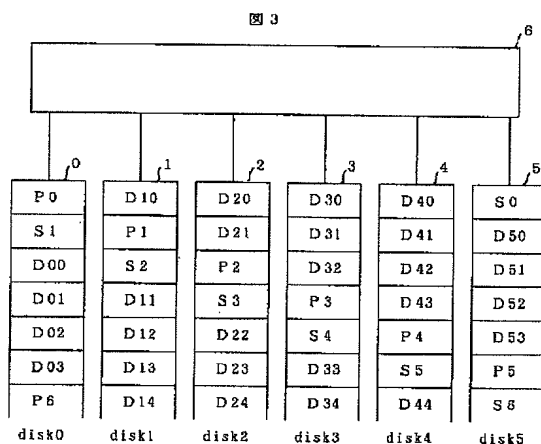
【図1】



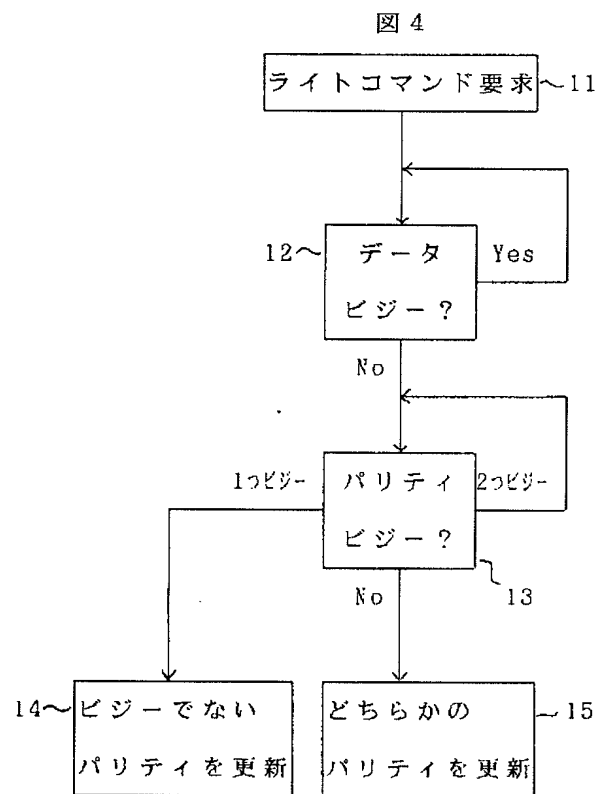
【図2】



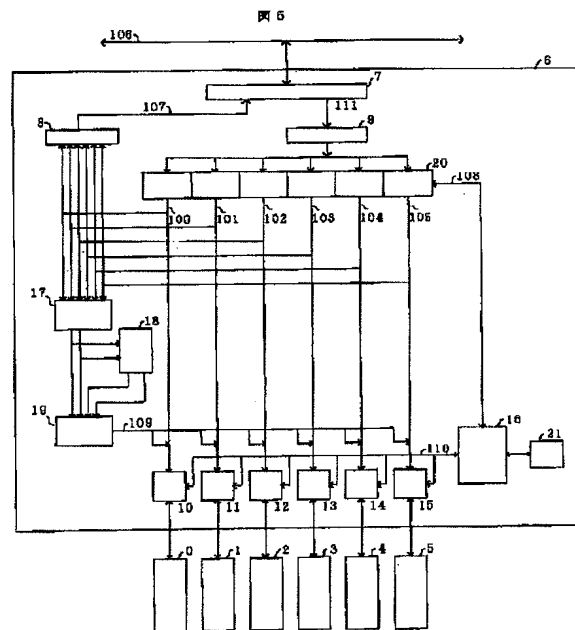
【図3】



【図4】



【図5】



フロントページの続き

(72)発明者 深谷 寧克  
愛知県尾張旭市晴丘町池上1番地株式会社  
日立製作所オフィスシステム事業部内

(72)発明者 福本 聡  
愛知県豊田市八草町八千草1247